CS 327E Milestone 9 due Sunday, 04/19.

Part 1:

1. Find a second dataset (aka `dataset2`) in CSV format that meets our dataset requirements.
2. Add a description of your dataset to the existing `DATASETS.txt` file.
3. Create a new folder in your Cloud Storage bucket and upload the dataset files to this folder.

Part 2:

Create a new Jupyter notebook called `<source>_ingest.ipynb` where `<source>` refers to the source of your secondary dataset (e.g. fda, bls, etc.). Implement the following logic in your notebook:

1. Create a new dataset in BQ for storing the staging tables for `dataset2`. The dataset should be named `<source>_staging`.
2. Import the dataset files from GCS into your new dataset in BQ. Ensure that you import each file into its own table.
3. Visually inspect the contents of each table by selecting a few sample records.

Part 3:

Create a new Jupyter notebook called `<source>_modeled.ipynb` where `<source>` refers to the source of your secondary dataset (e.g. fda, bls, etc.). Implement the following logic in your notebook:

1. Create a new dataset in BQ for storing the modeled tables. The modeled dataset should be named `<source>_modeled`.
2. Create modeled tables by applying the design principles from Milestone 4.
3. Each modeled table should have a primary key. Check for any primary key violations and deduplicate the records in SQL, if possible. Otherwise, make a note in `TRANSFORMS.txt` if there is a table which doesn't contain a valid primary key. You will need to deduplicate this table in the next milestone.
4. Check for any referential integrity violations between any parent and child tables.

Part 4:

1. Update your ERD to include the modeled tables in `dataset2`. Be sure to denote in the diagram the relationships between the tables within `dataset2` as well as **across** the two datasets if any exist. Name the ERD file `<source>_erd_modeled.pdf` where `<source>` is the source of your `dataset1`.

2. Think of 3 interesting queries that span your primary and secondary datasets. These queries should use a join or union or filter to combine the data from `dataset1` and `dataset2`. In addition, these queries should require some prior data transformation process to cleanse or standardize the data. These transformations will be done as part of the next milestone.

   For each of your 3 queries:

   - Briefly describe the expected results from the query and what SQL operations the query will use to produce those results (1-2 sentences).
   - Briefly describe what type(s) of data transforms are required to successfully implement the query (1-2 sentences).

   Create a file `CROSS-DATASETS.txt` and add your descriptions and explanations to this file.

CS 327E Milestone 9 Rubric

Due Date: 04/19/20

Part 1 - Edit the file `./DATASETS.txt` to include information on your secondary dataset. **-10** no description of secondary dataset in `DATASETS.txt`	10
Part 2 - Create a Jupyter notebook `<source>_ingest.ipynb` containing the ingestion pipeline as described in the outline. **-30** `<source>_ingest.ipynb` is missing. **-20** dataset `<source>_staging` not present in BQ project **-10** each missing staging table in BQ project **-10** inconsistent naming conventions across tables **-10** each missing sample records query from table	30
Part 3 - Create a Jupyter notebook `<source>_modeled.ipynb` containing the modeling pipeline as described in the outline. **-40** `<source>_modeled.ipynb` is missing. **-30** dataset `<source>_modeled` not present **-15** each missing modeled table, up to **-30** **-10** inconsistent naming conventions across tables **-10** each non-merged entity type, table with multiple entity types, or un-unioned tables containing the same data (i.e tables representing the same data across different years). **-10** each string field in modeled tables containing only `INTEGER`, `NUMERIC`, `DATE`, or `TIMESTAMP` not cast, up to **-30** **-10** each missing or incorrect primary key check and no valid explanation in `TRANSFORMS.txt` up to **-20** **-5** each missing or incorrect foreign key check and no valid explanation in `TRANSFORMS.txt` up to **-20**	40
Part 4 - Create a new ERD called `<source>_erd_modeled.pdf` which includes modeled tables from secondary dataset. Diagram their relationships as you have in previous milestones - this does include adding potential relationships between tables across both datasets. **-10** `<source>_erd_modeled.pdf` is missing. **-5** each incorrectly labeled key **-5** each incorrectly labeled relationship **-5** each incorrectly labeled data type Create a file `./CROSS-DATASETS.txt` containing query and transformation information for 3 queries, as described in the outline. Keep in mind that you do not actually have to *write* the query, just a description of one and transformations required to make the query work. **-10** `./CROSS-DATASETS.txt` does not exist **-5** for each missing pair of query description and required transformation(s) description, up to **-10**	20

`submission.json` submitted into Canvas. Your project **will not** be graded without this submission. The file should have the following schema:
`{`
`    "commit-id": "your most recent commit ID from Github",`
`    "project-id": "your project ID from GCP"`
`}`
Example:
`{`
`    "commit-id": "dab96492ac7d906368ac9c7a17cb0dbd670923d9",`
`    "project-id": "some-project-id"`
`}`
	Required

Total Credit:	**100**